First Capstone Project  
IMDB Movie Score Recommendation

June 13, 2017

# Overview

## Project Background and Description

How can we tell the greatness of a movie before it is released in cinema? This question puzzled almost everybody for a long time since there is no universal way to claim the goodness of movies. Many people rely on critics to gauge the quality of a film, while others use their instincts. But it takes the time to obtain a reasonable amount of critics review after a movie is released. And human instinct sometimes is unreliable.

Given that thousands of movies were produced each year, is there a better way for us to tell the greatness of movie without relying on critics or our own instincts?

## Project Scope

Predicting IMDB Score or greatness of a movie worldwide before it released in cinemas is my primary goal for this capstone project without relying on critics review data and human instincts data. My customer will be all cinema lovers or cinema producers/directors who can get a high level of overview of the probable score of the new movie they are releasing like a pre poll forecast.

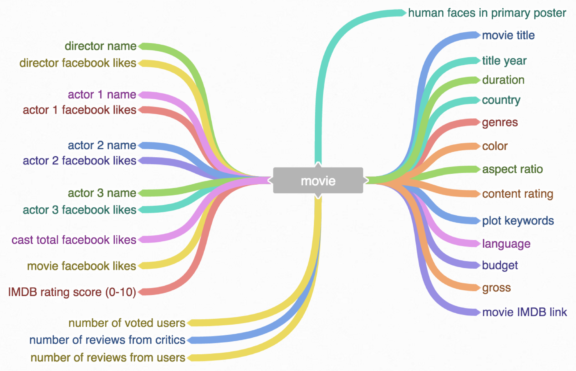
In a nutshell, I will be using different approaches to analyzing the IMDB dataset and then implementing some regression or classification model to provide visualizations of any related data I find in the dataset. This notebook will purely be an exploratory and hopefully concise enough attempt to explain the idea as well as using different methods to extract meaningful relations out of it.

My stakeholders will be greatly benefitted with my model and also this predicted model can bring more customers for the probable high IMDB score cinemas and will ultimately benefit the film producers/directors.

## High-Level Requirements

The image below shows all the 28 variables that I scraped from IMDB site. Roughly speaking, half of the variables is directly related to movies themselves, such as title, year, duration, etc. Another half is related to the people who involved in the production of the movies, eg, director names, director Facebook popularity, movie rating from critics, etc.

I will build a Machine Learning Model which will predict IMDB score of any movies before released in cinemas based on some input variable behavior and their relation with output predicted variable IMDB Score.

[](http://blog.nycdatascience.com/wp-content/uploads/2016/08/Screen-Shot-2016-08-21-at-1.09.28-PM.png)

I scraped 5000+ movies from IMDB website using the below link.

The scraping process took 2 hours to finish. In the end, I was able to obtain all needed 28 variables for 5043 movies and 4906 posters (998MB), spanning across 100 years in 66 countries. There are 2399 unique director names, and thousands of actors/actresses. Below are the 28 variables:

"movie title" "color" "num\_critic\_for\_reviews" "movie\_facebook\_likes" "duration" "director\_name" "director\_facebook\_likes" "actor\_3\_name" "actor\_3\_facebook\_likes" "actor\_2\_name" "actor\_2\_facebook\_likes" "actor\_1\_name" "actor\_1\_facebook\_likes" "gross" "genres" "num\_voted\_users" "cast\_total\_facebook\_likes" "facenumber\_in\_poster" "plot\_keywords" "movie\_imdb\_link" "num\_user\_for\_reviews" "language" "country" "content\_rating" "budget" "title\_year" "imdb\_score" "aspect\_ratio".

My solution to this will be like below and will change in future based on different scenarios:

* Filtering the Dataset to remove Null values and get only numerical columns. Standardizing the features.
* Create some kind of regression model based on certain input features say director\_facebook\_likes, actor\_1/2/3\_facebook\_likes, duration of the movie, budget of the movie, movie\_facebook\_likes.
* Train that model by 70% of the data from the dataset.
* Test that model by rest of the 30% data and start predicting.
* Create data visualizations/data story based on certain obvious facts for presentation.

## Deliverables

* Clear, documented code that will reproduce the entire analysis, and show what steps were taken in the analysis. Likely in a jupyter notebook format. Longer code could be provided separately and referenced in notebook.
* A slide deck presenting the results and implications of the analysis to the client.
* All deliverables will be in a GitHub repository.

<https://github.com/chatkausik/First-Capstone-Project>